AURESYS 2.0
Un Agent Intelligent au Service de l'Information Stratégique

Mannina Bruno, Quoniam Luc, Dou Henri

CRRM
Centre de Recherche Rétrospective de Marseille
Université Aix-Marseille III
13397 Marseille Cedex 20
Tel : 04-91-28-87-40, Fax : 04-91-28-87-12
mannina@crrm.univ-mrs.fr

Mots Clés : Agent Intelligent, Recherche, Information, Analyse Automatique

Résumé

Internet est aujourd'hui une source d'information impossible à négliger: Cinquante millions d'ordinateurs y sont connectés, qui peuvent offrir des bases de données ainsi que des sources d'informations informelles.
Au regard de cet énorme flux d'informations, il est primordial d'y rechercher et collecter de l'information sans s'éloigner du sujet à traiter et tout en étant suffisamment exhaustif.
Devant la taille de ce réseau, certains outils permettant aux utilisateurs de circuler sur les Inforoutes, doivent être considérés de nos jours comme incontournables. Ces outils sont appelés **Agents Intelligents**, ou plus couramment Robots.
Les agents intelligents sont des programmes capables de réagir avec un environnement, de s'adapter aux circonstances, de prendre des décisions ou d'enrichir eux-mêmes leur propre comportement, sur la base d'observations qu'ils effectuent. Les progrès en la matière sont très rapide, cependant peu de robots présentent toutes les caractéristiques suffisantes pour des applications en **veille technologique** et/ou en **analyse automatique**. Certains moteurs, tel que Lycos, Yahoo, etc... ne sont généralement pas très adéquate pour ce travail car ils ne sont pas paramétrables avec la souplesse nécessaire.

Auresys est un agent intelligent développé au CRRM qui permet à ses utilisateurs de parcourir l'ensemble des Inforoutes en recherchant les informations conformes aux besoins stratégiques de l'utilisateur. Auresys, par rapport aux différents agents existant sur l'**Internet**, se positionne comme un des rares moteurs intelligents répondant aux préoccupations nécessaires pour l'activité. Nous présenterons un comparatif des différents agents existants sur les inforoutes et leurs différentes fonctions, avec le positionnement concurrentiel d'Auresys.
Nous montrerons que cette méthode de recherche d'information permet un gain de temps et d'argent.
Nous démontrerons qu'à partir d'une récolte d'information depuis l'Internet, il est possible de créer des bases de données personnalisées et consultables à distance. A partir de ces bases de données personnalisées, nous montrerons aussi qu'il est possible d'affiner son interrogation en local afin d'avoir une information plus ciblée. Nous présenterons des exemples de traitements automatiques de l'information à but infostratégique.
Nous finirons en articulant cette source d'information et son traitement dans un processus de veille en développant le caractère indispensable de cette source d'information.

INTRODUCTION

Internet, plus qu'un réseau, est une interconnexion de réseaux, nationaux comme en France RENATER, ou régionaux comme aux Etats-Unis.
C'est en effet des Etats-Unis qu'a commencé à se développer Internet il y a 20 ans, à partir d'un réseau destiné à l'armée. Cette volonté de créer une interconnexion de petits réseaux et non un giganstesque réseau partageant toute l'information répondait à des impératifs stratégiques : en cas de destruction partielle, il fallait que le reste fonctionne.
Internet a naturellement perdu sa vocation militaire mais il a gardé cette organisation, ce qui permet une grande souplesse dans son évolution. Son organisme est donc hiérarchique : ainsi chaque ordinateur relié à Internet est en fait connecté à un réseau plus petit, lui-même connecté à un réseau plus grand... et chaque réseau s'interconnectant en une immense chaîne, ils constituent Internet.

Les agents intelligents permettent d'aider les utilisateurs à circuler sur les autoroutes de l'information en les soulageant de certaines tâches répétitives.
Ceux sont des programmes capables de réagir avec un environnement, de s'adapter aux circonstances, de prendre des décisions ou d'enrichir eux-mêmes leur propre comportement, sur la base d'observations qu'ils effectuent.

Le terme "Agent" est employé actuellement pour désigner un Robot ou un Logiciel capable d'assister un utilisateur dans la réalisation de tâches répétitives.
Dans le langage courant, le terme Agent est remplacé par celui de Robot.
Le vocabulaire est cependant très varié, puisque dans la littérature, plusieurs autres termes ont la même signification qu'Agent :

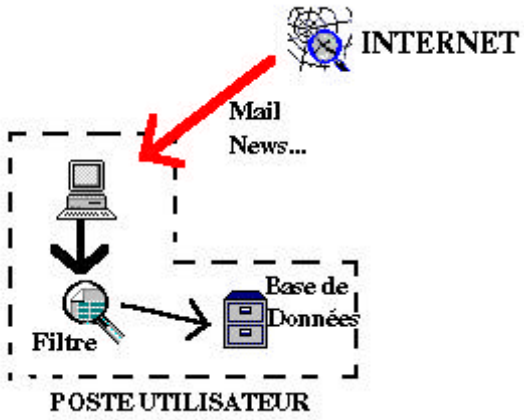
Robot, Spider, Wanderer, Web Worm, Web Walker, Moteur de Recherche, Bots, Brokers

Le terme "Intelligent" peut être remplacé par d'autres adjectifs qui définissent les agents plus précisement.

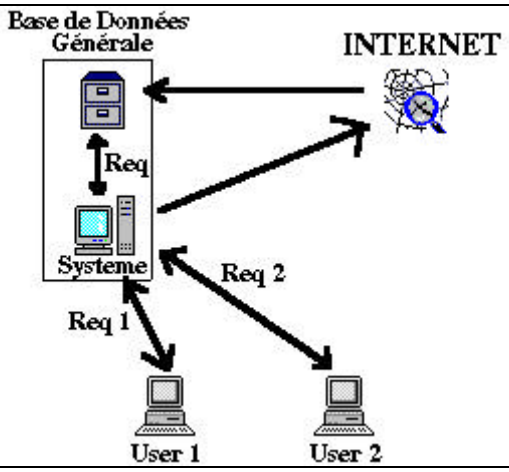
Un Agent Intelligent est capable d'apprendre, de se déplacer et de récupérer les informations les plus pertinentes.
Un Agent Personnalisé est capable de s'informer sur les habitudes de l'utilisateur.
Un Agent Flexible est capable de prendre des initiatives, et d'offrir des suggestions.
Un Agent Autonome est capable d'envoyer et de recevoir des informations sur son environnement immédiat.
Un Agent Spécialisé est capable de récupérer des informations sur des champs bien précis.

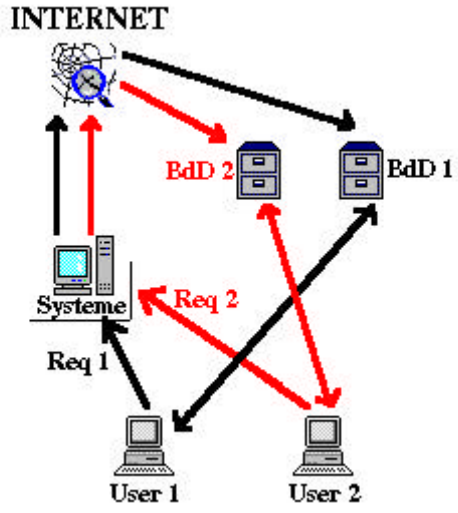
En ce qui concerne leur utilité, les robots permettent de récupérer des informations provenant d'Internet, de résoudre le problème de l'ingérable quantité d'information à laquelle est confronté l'utilisateur, de trouver des informations qui ne sont pas indexées (non présentes dans les serveurs conventionnels) de façon à avoir un avantage concurrentiel, d'économiser du temps pour le balayage des serveurs, de collecter automatiquement de l'information, d'engager une communication avec l'utilisateur ou même d'autres agents, de trouver les habitudes d'une personne, d'executer des taches lorsque l'utilisateur n'est pas présent.

Programmé généralement en Perl et/ou en C++, leur fonctionnement est variable. La figure 1 nous aidera à mieux comprendre les différents types d'agents existants.

<table>
<tr><th colspan="2">Robots Passifs</th></tr>
<tr><td> 
</td><td>Tous les utilisateurs qui sont connectés au Net, s'inscrivent généralement à des listes de discussion, reçoivent des messages (MAILS) et/ou s'abonnent à des NEWS.

Un robot passif à pour fonction de filtrer ces informations pour garder uniquement celles qui intéressent l'utilisateur.

Ces informations sont ensuite stockées dans des Bases de Données.</td></tr>
<tr><th colspan="2">Robots Actifs</th></tr>
<tr><td> 
</td><td>L'utilisateur, dans le cadre d'une recherche d'information envoie à un système spécifique une requête comportant des mots-clés.
Ce système, ayant auparavant indexés la plupart des informations présentes sur le NET, va chercher dans ses propres Bases de Données les informations répondant à la requête de l'utilisateur.

A aucun moment, l'utilisateur n'a accès directement aux Bases de Données.</td></tr>
<tr><th colspan="2">Robots Actifs Spécifiques</th></tr>
<tr><td> 
</td><td>A la différence des robots actifs, cette sorte de robot crée une Base de Données personnelle à l'utilisateur.
La réponse à une requête correspond au contenu d'une Base.

L'utilisateur pourra directement interroger la Base ainsi créée.</td></tr>
</table>

Figure 1 : Les différents types de Robots

1.4 Dangers

1.4.1 Hyper spécialisation

En affinant par étapes successives le type d'information que recherche un utilisateur, et en étudiant son comportement, pour en déduire les points récurrents, les agents risquent d'augmenter la précision et la spécialisation de l'information recherchée.
Une sorte de spirale infernale, qui pourrait conduire à ne plus recevoir qu'un type d'information unique.
Alors qu'il est primordial pour un chercheur de se repositionner constamment dans son domaine, les robots ont tendance à réduire le champ de vision de celui-ci.
Un Aller-Retour entre le domaine étudié et son environnement est nécessaire pour une meilleure évaluation des recherches.
Ainsi, un chercheur aura la possibilité d'appliquer ses recherches à d'autres domaines s'il est suffisamment au courant de ce qui s'y fait.

1.4.2 Passivité de l'utilisateur

Par essence, Internet n'est pas un support de diffusion de l'information, au même titre que la radio ou la télévision par exemple. Une bonne partie de l'intérêt de NET réside dans le fait que ses utilisateurs vont eux-mêmes chercher l'information qu'ils souhaitent, et ne se contentent plus de recevoir cette information.
Finalement, la généralisation des agents, mêmes intelligents, pourrait transformer le NET en une sorte de support redevenu "à sens unique".

1.4.3 Contrôle de l'information par la machine

La crainte (classique, certes) de voir les machines s'approprier une partie des prérogatives humaines peut sembler justifiée. En accordant aux agents l'autonomie qu'ils "réclament", les humains risquent de leur accorder une importance qui les dépassent.
Et en décidant des informations transmises, les agents vont exercer une sorte de contrôle, amoindrissant ainsi l'exercice de notre esprit critique.

Cette inquiétude, relevée dans la presse on-line (Cybersphère), est quelque peu pessimiste. En effet, notre société n'est pas encore arrivée à l'ère où la machine prendra le pas sur l'homme (comme dans certains films de science fiction). Notre esprit critique est sauf dans la mesure où la machine respecte les requêtes qui lui sont paramètrées et n'a pas la possibilités de les modifier.

1.5 Comment réagir à la visite d'un robot ?

Pour détecter la visite d'un robot sur un serveur, il faut regarder le fichier "message" du serveur. Il comporte des informations sur les visiteurs du site.
Les robots ont, dans la plupart des cas, une signature particulière : Nom/version.
Dans le cas d'AURESYS, sa signature est : User-Agent : auresys/1.0

1.5.1 Méthodes

✓ Ne pas être surpris : Les robots sont seulement des programmes qui visitent les sites WWW pour les indexer. C'est très utile si l'on veut que son serveur se fasse connaître, ce qui est généralement le but recherché.

✓ Ne pas réagir violemment : Il est inutile de renvoyer des tonnes d'informations en pensant que le robot sera submergé, car celui-ci utilise généralement des procédures spéciales qui lui permettront d'ignorer une trop grosse quantité d'information.

✓ Regarder les "logs" du serveur : Vérifier ce qu'ont fait exactement les robots et quelles informations ils ont récupéré. Les champs HTTP **From** et **User-Agent** renseignent sur la personne responsable du robot.

✓ Trouver des informations sur le robot : Il existe sur le NET une liste des robots actifs : http://info.webcrawler.com/mak/projects/robots/active.html

✓ En savoir plus : Si aucun détail n'a été trouvé sur la personne qui a lancé le robot, il faut envoyer un message à : **comp.infosystems.www.providers** du USENET. Il est possible qu'une autre personne ait eu le même problème, ou que la personne ayant lancé le robot soit à l'écoute.

✓ Etre constructif : Si un contact s'engage avec la personne qui a lancé le robot, il faut lui indiquer les aspects gênants de son robot. Ces informations lui seront sans doute très utile et rentrent totalement dans l'esprit d'Internet.

✓ Partager son expérience : Ne pas hésiter à diffuser des informations supplémentaires concernant les robots déjà rencontrés.

1.5.2 Interdir et autoriser l’accés à un Agent.

Il existe à l’heure actuelle deux méthodes pour interdir aux agents d’accéder à un serveur, mais aucune d’elles n’est un moyen éfficace car toutes deux ne sont utiles que si le propriétaire de l’agent a pris le soin d’inclure la procédure permettante de respecter l’accés.

1.5.2.1 Le fichier "robot.txt"

Tous les robots doivent respecter un protocole d'exclusion. Celui-ci consiste en la désignation des répertoires qui peuvent être visités. Chaque robot, avant toute visite de site, doit parcourir ce fichier.

```
# /robots.txt for http://crrm.univ-mrs.fr/

User-Agent: Lycos        # Match Lycos
Disallow: /spool

User-Agent: *  # Match any robot
Disallow: /tmp  /spool
```

Figure 1 : Exemple de fichier robot.txt

Cette figure représente le listing du fichier d'exclusion des robots.
Le robot Lycos a accès à tous les répertoire, sauf "/spool".
Tous les autres robots ne pourront pas visiter les répertoires "/spool" et "/tmp".

1.5.2.2 Le Robot Meta Tag des pages HTML

Le Robot Meta Tag est une simple commande qui indique aux agents si ils sont autorisés à indexer la page HTML et si ils sont peuvent utiliser les différents liens existants dans cette même page.
Note : Actuellement peu d'Agents respectent ce Tag.

Comme tous les Meta Tag, il doit être placé dans le HEAD de la page HTML.

```
<HTML>
<HEAD>
<meta name="robots" content="noindex, nofollow">
<meta name="description" content="Cette page est ....">
<TITLE> ... Les Agents Intelligents ... </TITLE>
</HEAD>
<BODY>
...
```

Figure 2 : Le Robot Meta Tag

Voici quelques exemples pour bien comprendre comment l'utiliser :

```
#Permet l'indexation et l'utilisation des autres liens
<meta name="robots" content="index, follow">

#Ne permet pas l'indexation de la page, mais autorise l'agent à utiliser les liens hypertextes pour poursuivre sa recherche
<meta name="robots" content="noindex, follow">

#Permet l'indexation de la page, mais l'agent ne doit pas utilser les liens de la page pour continuer sa recherche
<meta name="robots" content="index, nofollow">

#N'autorise n'y l'indexation de la page, n'y l'utilisation des liens hypertexte de la page HTML
<meta name="robots" content="noindex, nofollow">
```

Figure 3 : Exemples d'utilisation du Meta Tag Robot

2. Présentation d'AURESYS

AURESYS = AUtomated REsearch SYStem
= Système de Recherche Automatisée

2.1 Ce qu'il fait

AURESYS permet de créer des Bases de Données interrogeables à distance par plusieurs utilisateurs.
Les informations sont récupérées du NET, puis traitées pour être stockées dans ces bases.
Chaque Base de Données répond à une **requête personnalisée** d'un utilisateur.

2.2 Son utilité

Grâce aux Bases de Données crées, AURESYS donne les moyens à un utilisateur d'avoir un **maximum de renseignements** se rapportant à un sujet donné.

L'utilisateur peut constituer un **corpus analysable** directement par des outils bibliomètriques, ainsi que des **Dossiers Généraux d'Information** (DGI) ou des **Dossiers d'Information Stratégique** (DIS).

2.3 Ses atouts

- AURESYS respecte le protocole d'exclusion des robots. Seuls les répertoires qui sont autorisés sont visités (robot.txt : cf. 5.2.2).
- Les bases de Données crées sont **réactualisables**, d'où son intérêt en **Veille Technologique.**
- AURESYS permet de **choisir le domaine** dans lequel le robot effectuera sa recherche, ainsi que **la profondeur de recherche** (liens hypertexte).
- Possibilité de commencer la recherche depuis **n'importe quelle page** d'un site.

■ Grâce à **l'incrémentation des numéros IP**, AURESYS est capable de trouver des sites qui ne sont pas enregistrés par les utilitaires classiques du WEB (AltaVista, Yahoo, Lycos, Search...), la plupart des autres robots se contentant d'interroger les sites déjà indexés.

■

2.4 Ses utilisateurs

Toute personne désirant s'informer sur **un sujet donné avec réactualisation permanente** des informations.
(Veilleur Technologique ou Concurrentiel par exemple)

Exemple : J'ai réalisé pour un étudiant de DEA, désirant récupérer le plus d'informations possible sur les vins, une Base de Données regroupant uniquement les informations qui correspondent à sa requête. Cet étudiant désirait connaître la production, l'exportation et la vente des vins dans le monde.
La Base de Données constituée comprend plus de 250 références directement utilisables par des outils bibliométriques (Dataview).

2.5 Sa conception

- Programmation d'AURESYS : **Langage PERL5** : Ce choix est du à la diversité des fonctions système. De plus, c'est le langage le plus courant dans la programmation des robots.

- Interface INTERNET : **Langage HTML** : Standard d'Internet (seul langage possible, le format PDF n'ayant pas d'interface avec l'utilisateur).

- Interface HTML AURESYS : **Langage Shell** : Grâce à ce langage, il est possible de récupérer les paramétres du robot depuis la page HTML de configuration.

- Utilisation de **FREEWAIS** pour indexer la Base de Données : Freewais est un programme en FreeWare, qui permet de créer des Bases de Données.

- Utilisation de **SFGATE** pour interroger la Base de Données : Ce programme permet

Pour plus d'informations sur Internet :

Un nouveau guide Internet : http://www.imaginet.fr/~gmaire/manuel.htm
Ce guide est très complet (réseau, Nestcape, ftp, Telnet, tous les langages et les utilitaires liés au NET...).

Langage/Utilitaire	Références
PERL5	http://www.perl.com
HTML	**A Beginner's Guide to HTML** http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html **Style Guide for Online Hypertext** http://www.w3.org/hypertext/WWW/Provider/Style/All.html
Shell	
FreeWais	http://ls6-www.informatik.uni-dortmund.de
SFGATE	http://ls6-www.informatik.uni-dortmund.de/Sfgate

2.6 Ses améliorations futures

- Rapidité : -de connexion.
  - de traitement de l'information.
- Fléxibilité d'utilisation
- Evaluation d'un critère de multimédia.
- Paramétrage des mots-clés.
- Association d'un lemmatiseur automatique d'information.

BIBLIOGRAPHIE :